OneLLM: One Framework to Align All Modalities with Language

Jiaming Han[1,2], Kaixiong Gong[1,2], Yiyuan Zhang[1,2], Jiaqi Wang[2], Kaipeng Zhang[2]
Dahua Lin[1,2], Yu Qiao[2], Peng Gao[2], Xiangyu Yue[1†]

[1]MMLab, The Chinese University of Hong Kong
[2]Shanghai Artificial Intelligence Laboratory

Abstract

*Multimodal large language models (MLLMs) have gained significant attention due to their strong multimodal understanding capability. However, existing works rely heavily on modality-specific encoders, which usually differ in architecture and are limited to common modalities. In this paper, we present **OneLLM**, an MLLM that aligns **eight** modalities to language using a unified framework. We achieve this through a unified multimodal encoder and a progressive multimodal alignment pipeline. In detail, we first train an image projection module to connect a vision encoder with LLM. Then, we build a universal projection module (UPM) by mixing multiple image projection modules and dynamic routing. Finally, we progressively align more modalities to LLM with the UPM. To fully leverage the potential of OneLLM in following instructions, we also curated a comprehensive multimodal instruction dataset, including **2M** items from image, audio, video, point cloud, depth/normal map, IMU and fMRI brain activity. OneLLM is evaluated on **25** diverse benchmarks, encompassing tasks such as multimodal captioning, question answering and reasoning, where it delivers excellent performance. Code, data, model and online demo will be available at* https://github.com/csuhan/OneLLM.

1. Introduction

Large Language Models (LLMs) are getting increasingly popular in the research community and industry due to their powerful language understanding and reasoning capabilities. Notably, LLMs such as GPT4 [62] have reached performance nearly on par with humans in various academic exams. The progress in LLMs has also inspired researchers to employ LLMs as an interface for multimodal tasks, such as vision-language learning [4, 44], audio and speech recognition [25, 99], video understanding [11, 45, 100], *etc*.
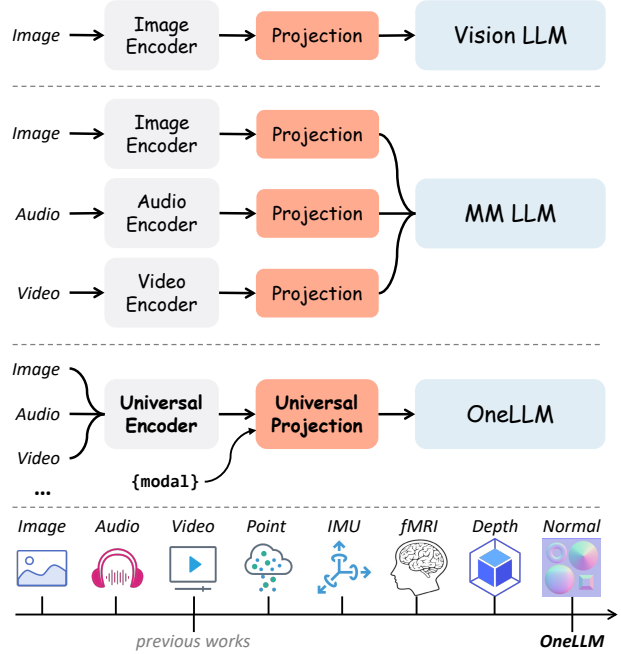

Figure 1. **Comparisons of Different Multimodal LLMs.** Vision LLM: one image encoder and projection module. Multimodal (MM) LLM: modality-specific encoder and projection module. **OneLLM**: a universal encoder, a universal projection module and modality tokens {modal} to switch between modalities. **Bottom:** OneLLM expands supported modalities from three to eight.

Among these tasks, vision-language learning is the most active field, with more than 50 vision LLMs proposed in the recent half-year alone [20]. Typically, a vision LLM comprises a visual encoder, an LLM, and a projection module connecting the two components. The vision LLM is first trained on massive paired image-text data [70] for vision-language alignment and then fine-tuned on visual instruction datasets, enabling it to complete various instructions tied to visual inputs. Beyond vision, significant efforts have been invested in developing other modality-specific LLMs, such as audio [25], video [45], and point clouds [28]. These models generally mirror the architectural framework

---

1

and training methodology of vision LLMs, and rely on the solid foundation of pretrained modality-specific encoders and well-curated instruction-tuning datasets for their effectiveness.

There are also several attempts to integrate multiple modalities into one MLLM [10, 31, 59, 104]. As an extension of vision LLM, most previous works align each modality with the LLM using modality-specific encoders and projection modules (middle of Fig. 1). For instance, X-LLM [10] and ChatBridge [104] connect pretrained image, video, and audio encoders with LLMs using separate Q-Former [44] or Perceiver [35] models. However, these modality-specific encoders usually differ in architecture and considerable effort is required to unify them into a single framework. Furthermore, pretrained encoders that deliver reliable performance are usually restricted to widely used modalities such as image, audio, and video. This limitation poses a constraint on MLLMs' ability to expand to more modalities. Thus, a crucial challenge for MLLMs is *how to build a unified and scalable encoder capable of handling a wide range of modalities.*

We get inspiration from recent works on transferring pretrained transformers to downstream modalities [51, 57, 88, 103]. Lu *et al.* [51] proved that a frozen language-pretrained transformer can achieve strong performance on downstream modalities such as image classification. Meta-Transformer [103] demonstrated that a frozen visual encoder can achieve competitive results across 12 different data modalities. The insights from the works mentioned above suggest that pretrained encoders for each modality may not be necessary. Instead, a well-pretrained transformer may serve as a universal cross-modal encoder.

In this paper, we present **OneLLM**, an MLLM that aligns eight modalities to language using one unified framework. As shown in Fig. 1, OneLLM consists of lightweight modality tokenizers, a universal encoder, a universal projection module (UPM), and an LLM. In contrast to prior works, the encoder and projection module in OneLLM are shared across all modalities. The modality-specific tokenizers, each comprised of only one convolution layer, convert input signals into a sequence of tokens. Additionally, we add learnable modality tokens to enable modality switching and transform input tokens of diverse lengths into tokens of a fixed length.

Training a model of this complexity from scratch poses significant challenges. We start from a vision LLM and align other modalities to the LLM in a progressive way. In detail, **(i)** we build a vision LLM with pretrained CLIP-ViT [67] as the image encoder, accompanied by several transformer layers as the image projection module, and LLaMA2 [78] as the LLM. After pretraining on massive paired image-text data, the projection module learns to map visual representations into the embedding space of LLM.

(ii) To align with more modalities, we need a universal encoder and projection module. As discussed before, the pretrained CLIP-ViT is possible to serve as a universal encoder. For UPM, we propose to mix multiple image projection experts as a universal X-to-language interface. To increase the model capability, we also design a dynamic router to control the weight of each expert for the given inputs, which turns UPM into soft mixtures-of-experts [66]. Finally, we progressively align more modalities wiht the LLM based on their data magnitude.

We also curate a large-scale multimodal instruction dataset, including captioning, question answering, and reasoning tasks across eight modalities such as image, audio, video, point clouds, depth/normal map, Inertial Measurement Unit (IMU), and functional Magnetic Resonance Imaging (fMRI). By finetuning on this dataset, OneLLM has strong multimodal understanding, reasoning, and instruction-following capabilities. We evaluate OneLLM on multimodal captioning, question answering and reasoning benchmarks where it achieves superior performance than previous specialized models and MLLMs. In conclusion, we summary our contributions as:

- We propose a unified framework to align multimodal inputs with language. Different from existing works with modality-specific encoders, we show that a unified multimodal encoder, which leverages a pretrained vision-language model and a mixture of projection experts, can serve as a general and scalable component for MLLMs.
- To the best of our knowledge, OneLLM is the first MLLM that integrates eight distinct modalities within a single model. With the unified framework and progressive multimodal alignment pipeline, OneLLM can be easily extended to incorporate more data modalities.
- We curate a large-scale multimodal instruction dataset. OneLLM finetuned on this dataset achieves superior performance on multimodal tasks, outperforming both specialist models and existing MLLMs.

2. Related Work

Large Vision-Language Models. Large Language Models (LLMs) have gained a lot of attention recently. Therefore, extending LLMs to the vision domain is an emergent and rapidly growing research area. Flamingo [4] is a pioneer to inject frozen visual features into LLM with cross-attention layers, achieving superior performance on a wide range of vision-language tasks. BLIP2 [44] uses a Q-Former to aggregate visual features into a few tokens aligned with LLM. Recently, with the popularity of instruction-following LLMs, vision LLMs have experienced a new explosion. LLaMA-Adapter [21, 102] connect pretrained CLIP [67] and LLaMA [78] with parameter-efficient fine-tuning methods, which can tackle close-set visual question answering and image captioning tasks. Subsequent works [21, 48, 95,
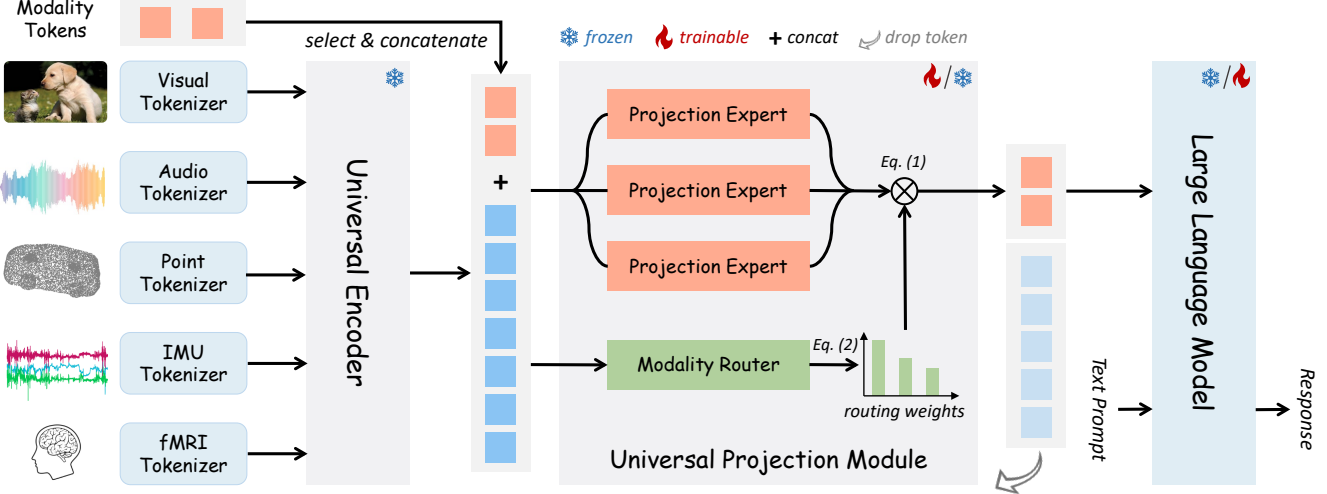
Figure 2. **The Architecture of OneLLM**, which consists of modality tokenizers, a universal encoder, a universal projection module (UPM) and an LLM. The modality tokenizer is a 2D/1D convolution layer to transform the input signal into a sequence of tokens. For simplicity, we omit video, depth/normal map tokenizers. The universal encoder is a frozen vision-language model (*i.e.* CLIP [67]) to extract high dimensional features. The UPM is composed of several projection experts and modality routers to align the input signal with language. For the alignment stage, we train modality tokenizers and UPM, and keep LLM frozen. For the instruction tuning stage, we only train the LLM and keep other models frozen. In a forward pass of UPM, we concatenate the input and modality tokens as input. Then we only take the modality tokens as a summary of the input signal and feed it into LLM for multimodal understanding.

105] propose to train such model on large-scale image-text data, enabling it to complete various instructions about images. Among them, LLaVA [48] adopt a linear layer to directly project visual tokens into LLMs, while MiniGPT-4 [105] and some other works [21, 95] resample visual tokens into fixed-length tokens, reducing the computation cost of LLMs. Our work also belongs to the later branch. We preset learnable tokens for each modality (*i.e.*, modality tokens), which are then used to aggregate input information and generate fixed-length tokens for all modalities.

Multimodal Large Language Models. In addition to vision LLMs, recent works proposed to extend LLMs to other modalities, such as audio [25, 99], video [11, 45, 100] and point cloud [28, 92]. These works make it possible to unify multiple modalities into one LLM. X-LLM [10] adopts modality-specific Q-Former [44] and adapters to connect pretrained image, audio and video encoders with LLMs. ChatBridge [104] and AnyMAL [59] follow a similar architecture with X-LLM but adopts Perceiver [35] and linear layers respectively to align modality encoders with LLMs. Meanwhile, PandaGPT [77] and ImageBind-LLM [31] utilize ImageBind [23] as the modality encoder and therefore naturally support multimodal inputs. However, current MLLMs are limited to supporting common modalities such as image, audio and video. It remains unclear how to expand MLLMs to more modalities with a unified framework. In this work, we propose a unified multimodal encoder to align all modalities with language. We show that one universal encoder and projection module can effectively map multimodal inputs to LLM. To our knowledge, OneLLM is

first MLLM capable of supporting eight distinct modalities.

Multimodal-Text Alignment. Aligning multiple modalities into one joint embedding space is important for cross-modal tasks, which can be divided into two lines of works: discriminative alignment and generative alignment. The most representative work of discriminative alignment is CLIP [67], which utilize contrastive learning to align image and text. Follow-up works extend CLIP to audio-text [30, 85], video-text [53, 90], point-text [101] *etc.* Besides, ImageBind [23] proposes to bind various modalities to images with contrastive learning. On the other hand, generative alignment has attracted much attention in the era of LLM. GIT [82] aligns image and text using a generative image-to-text transformer. BLIP2 [44] proposes generative pretraining to connect frozen vision encoder and LLM. VALOR [12] and VAST [13] extends the training paradigm of BLIP2 to more modalities such as audio and video. Our work also belongs to generative alignment. In contrast to prior works, we directly align mutlimodal inputs to LLMs, thus getting rid of the stage of training modality encoders.

3. Method

In this section, we will first introduce the architecture of OneLLM (Sec. 3.1) and then present our two training phases: progressive multimodal alignment (Sec. 3.2) and unified multimodal instruction tuning (Sec. 3.3).

3.1. Model Architecture

Fig. 2 depicts the four main components of OneLLM: modality-specific tokenizers, a universal encoder, a universal projection module (UPM) and an LLM. Detailed descriptions are presented in the following sections.

Lightweight Modality Tokenizers. The modality tokenizer is to transform the input signal into a sequence of tokens, thereby a transformer-based encoder can process these tokens. We denote the input tokens as $\mathbf{x} \in \mathbb{R}^{L \times D}$, where L is the sequence length and D is the token dimension. Considering the variations inherent to different data modalities, we design a separate tokenizer for each modality. For visual inputs with 2D position information such as image and video, we directly utilize a single 2D convolution layer as the tokenizer. For other modalities, we transform the input into a 2D or 1D sequence, which is then tokenized using a 2D/1D convolution layer. For example, we transform audio signals into 2D spectrogram and sample a subset of point clouds with 2D geometric prior. Due to space limit, please refer to Sec. C.1 of the appendix for more details.

Universal Encoder. As discussed in Sec. 1, frozen pretrained transformers demonstrate strong modality transfer capability [51, 103]. Therefore, we leverage pretrained vision-language models as the universal encoder for all modalities. Vision-language models, when trained on extensive image-text data, typically learn robust alignment between vision and language, so they can be easily transferred to other modalities. In OneLLM, we use CLIP-ViT [67] as a universal computation engine. Following previous works [51, 103], we keep the parameters of CLIP-ViT frozen during training. Note that for video signals, we will feed all video frames into the encoder in parallel and perform token-wise averaging between frames to speed up training. Other strategies, such as token concatenation, may further enhance the model's video understanding capability.

Universal Projection Module. In contrast to existing works with modality-specific projection, we propose a Universal Projection Module (UPM) to project any modality into LLM's embedding space. As shown in Fig. 2, UPM consists of K projection experts $\{P_k\}$, where each expert is a stack of transformer layers pretrained on image-text data (will discuss in Sec. 3.2). Although one expert can also realize any modality-to-LLM projection, our empirical findings suggest that multiple experts are more effective and scalable. When scaling to more modalities, we only need to add a few parallel experts.

To integrate multiple experts into one module, we propose a dynamic *modality router* R to control each expert's contribution and increase the model capacity. The router R is structured as a straightforward Multi-Layer Perception that receives input tokens and calculates the routing weights for each expert, *i.e.*, soft router [66]. We will also discuss other types of router in Sec. 4.3, such as constant router and

sparse router. Besides, we add learnable modality tokens $\{\mathbf{q}_m\}_{m \in \mathcal{M}}$ to switch between modalities, where $\mathcal{M}$ is the set of modalities and $\mathbf{q}_m \in \mathbb{R}^{N \times D}$ contains N tokens of dimension D. In a forward pass for modality m, we feed the concatenation of input tokens $\mathbf{x}_m \in \mathbb{R}^{L \times D}$ and modality tokens $\mathbf{q}_m$ into UPM:

$$[\bar{\mathbf{q}}_m, \bar{\mathbf{x}}_m] = \text{UPM}([\mathbf{q}_m, \mathbf{x}_m]) = \sum_{k=1}^{K} \mathbf{w}_m \cdot P_k([\mathbf{q}_m, \mathbf{x}_m]), \quad (1)$$

$$\mathbf{w}_m = \sigma \circ R_m([\mathbf{q}_m, \mathbf{x}_m]), \quad (2)$$

where $\mathbf{w}_m \in \mathbb{R}^{N \times K}$ is the routing weight and the SoftMax function σ is to ensure $\sum_{k=1}^{K} \mathbf{w}_{m,k} = 1$. For any modality m, we *only* extract the projected modality tokens $\bar{\mathbf{q}}_m$ as a summary of input signals, transforming $\mathbf{x}_m$ from varying lengths into uniform, fixed-length tokens.

LLM. We employ the open-source LLaMA2 [79] as the LLM in our framework. The input to LLM includes projected modality tokens $\bar{\mathbf{q}}_m$ and the text prompt after word embedding. Note we always put modality tokens at the beginning of the input sequence for simplicity. Then LLM is asked to generate appropriate response conditioned on modality tokens and text prompt.

3.2. Progressive Multimodal Alignment

Image-text alignment has been well investigated in previous works [21, 49, 105]. Therefore, a naive approach for multimodal alignment is to jointly train the model on multimodal-text data. However, training models directly on multimodal data can lead to biased representations between modalities due to the imbalance of data scale. Here we propose to train an image-to-text model as initialization and progressively ground other modalities into LLM.

Image-Text Alignment. We begin with a basic vision LLM framework, comprising an image tokenizer, a pretrained CLIP-ViT, an image projection module P_I and an LLM. Considering that image-text data is relatively abundant compared to other modalities, we first train the model on image-text data to well align CLIP-ViT and LLM, *i.e.*, learning a good image-to-text projection module. The pretrained P_I not only serves as a bridge connecting images and language, but also provides a good initialization for multimodal-text alignment. Then we build UPM by mixing multiple pretrained P_I: $\text{UPM} = \{P_k\}_{k=1}^{K} = \{P_I\}_{k=1}^{K}$, reducing the cost of aligning other modalities to language.

Multimodal-Text Alignment. We formulate multimodal-text alignment as a continual learning process [80]. At timestamp t, we have trained the model on a set of modalities $\mathcal{M}_1 \cup \mathcal{M}_2 \cdots \mathcal{M}_{t-1}$, and the current training data is from $\mathcal{M}_t$. To prevent catastrophic forgetting, we will sample balanced data from previous trained data and current data. In our case, we divide multimodal-text alignment into multiple training stages based on their data magnitude: stage I (image), stage II (video, audio and point cloud)

4

and stage III (depth/normal map, IMU and fMRI). If we want to support new modalities, we can repeat the training episode, *i.e.*, sampling a similar amount of data from previous modalities and jointly training the model with the current modalities.

Multimodal-Text Dataset. We collect X-text pairs for each modality. The image-text pairs include LAION-400M [70] and LAION-COCO [69]. The training data for video, audio and point clouds are WebVid-2.5M [8], WavCaps [56] and Cap3D [54], respectively. Since there is no large-scale depth/normal map-text data, we use pretrained DPT model [19, 68] to generate depth/normal map. The source images and text and from CC3M [73]. For IMU-text pairs, we use the IMU sensor data of Ego4D [27]. For fMRI-text pairs, we use fMRI signals from the NSD [5] dataset and take the captions associated with the visual stimuli as text annotations. Note that the input to LLM is the concatenation of modality tokens and caption tokens. We do not add system prompts at this stage to reduce the number of tokens and speed up training.

3.3. Unified Multimodal Instruction Tuning

After multimodal-text alignment, OneLLM becomes a multimodal captioning model which can generate a short description for any input. To fully unleash OneLLM's multimodal understanding and reasoning capabilities, we curate a large-scale multimodal instruction tuning dataset to further finetune OneLLM.

Multimodal Instruction Tuning Dataset. We collect instruction tuning (IT) dataset for each modality. Following previous works [15, 48], the *image* IT datasets are sampled from the following datasets: LLaVA-150K [49], COCO Caption [14], VQAv2 [26], GQA [34], OKVQA [55], A-OKVQA [71], OCRVQA [58], RefCOCO [36] and Visual Genome [38]. The *video* IT datasets include MSRVTT-Cap [91], MSRVTT-QA [89] and video instruction data from [104]. The *audio* IT datasets include AudioCaps [37] and audio conversation data from [104]. The *point cloud* IT dataset is a 70K point cloud description, conversation and reasoning dataset from [92]. The *depth/normal map* IT datasets are generated from image IT datasets: we random sample 50K visual instruction data from LLaVA-150K and generate depth/normal map using DPT model [19]. For *IMU* and *fMRI* IT datasets, we also random sample a subset from Ego4D [27] and NSD [5], respectively. Finally, our mutlimodal IT datasets have about **2M** items, covering multiple tasks such as detailed description/reasoning, conversation, short question answering and captioning.

Prompt Design. Given the diverse modalities and tasks within our multimodal IT datasets, we carefully design the prompts to avoid conflicts between them. **(a)** When utilizing IT datasets generated by GPT4 (*e.g.*, LLaVA-150K), we adopt the original prompts provided by these datasets. **(b)**

For captioning tasks, we empoly the prompt: *Provide a one-sentence caption for the provided {modal}*. **(c)** For open-ended question answering tasks, we enhance the question with *Answer the question using a single word or phrase*. **(d)** For question answering tasks with options, the prompt is: *{Question} {Options} Answer with the option's letter from the given choices directly*. **(e)** For IMU and fMRI datasets, we apply prompt such as *Describe the motion* and *Describe this scene based on fMRI data*. Despite using these fixed prompts, our experiments indicate that OneLLM is capable of generalizing to open-ended prompts during inference. For detailed prompts on each task and modality, please check out Sec. C.4 of the appendix.

In the instruction tuning stage, we organize the input sequence as: $\{\bar{q}, Sys, [Ins_t, Ans_t]_{t=1}^T\}$ where $\bar{q}$ is the modality tokens, Sys is the system prompt, $[Ins_t, Ans_t]$ corresponds to the t-th instruction-answer pair in a conversation. Note that for multimodal inputs involving multiple modalities, such as audio-visual tasks [42], we position all modality tokens at the start of the input sequence.

We fully finetune the LLM and keep rest parameters frozen. Although recent works often employ parameter-efficient methods [33], we empirically show that the full finetuning approach more effectively harnesses the multimodal capabilities of OneLLM, particularly with the utilization of smaller LLMs (*e.g.*, LLaMA2-7B).

4. Experiment

4.1. Implementation Details

Architecture. The universal encoder is CLIP VIT Large pretrained on LAION [70]. The LLM is LLaMA2-7B [79]. The UPM has K=3 projection experts, where each expert has eight Transformer blocks and 88M parameters. The size of modality tokens for each modality is $\mathbb{R}^{30 \times 1024}$.

Training Details. We use AdamW optimizer with β_1=0.9, β_2=0.95 and weight decay of 0.1. We apply a linear learning rate warmup during the first 2K iterations. For stage I, we train OneLLM on 16 A100 GPUs for 200K iterations. The effective batch size (using gradient accumulation) is 5120. The maximum learning rate is 5e-5. For stage II (*resp.* III), we train OneLLM on 8 GPUs for 200K (*resp.* 100K) with an effective batch size of 1080 and maximum learning rate of 1e-5. In the instruction tuning stage, we train OneLLM on 8 GPUs for 1 epoch (96K) with an effective batch size of 512 and maximum learning rate of 2e-5.

4.2. Quantitative Evaluation

We evaluate OneLLM on multimodal tasks and put evaluation details to Sec. D of the appendix.

Image-Text Evaluation. In Tab. 1, we evaluate OneLLM on visual question answering (VQA), image captioning and recent multimodal benchmarks. For VQA tasks, OneLLM-

Model	LLM	VQA						Image Caption		MM Benchmark			
		GQA	VQAv2	OKVQA	TVQA	SQA	Vizwiz	NoCaps	Flickr	MME	MMB	MMVet	SEED
vision specialized LLM													
Flamingo-9B [4]	Chinchilla-7B	-	51.8	44.7	30.1	-	28.8	-	61.5	-	-	-	-
Flamingo-80B [4]	Chinchilla-70B	-	56.3	50.6	31.8	-	31.6	-	67.2	-	-	-	-
BLIP-2 [44]	Vicuna-7B	-	-	-	40.1	53.8	-	107.5	74.9	-	-	-	-
BLIP-2 [44]	Vicuna-13B	41.0	41.0	-	42.5	61	19.6	103.9	71.6	1293.8	-	22.4	-
InstructBLIP [15]	Vicuna-7B	49.2	-	-	50.1	60.5	34.5	123.1	82.4	-	36	26.2	-
InstructBLIP [15]	Vicuna-13B	49.5	-	-	50.7	63.1	34.3	121.9	82.8	1212.8	-	25.6	-
IDEFICS-9B [39]	LLaMA-7B	38.4	50.9	38.4	25.9	-	35.5	-	27.3	-	48.2	-	-
IDEFICS-80B [39]	LLaMA-65B	45.2	60.0	45.2	30.9	-	36.0	-	53.7	-	54.5	-	-
LLaMA-Ad.v2 [21]	LLaMA-7B	43.9	-	55.9	43.8	54.2	-	42.7	30.5	972.7	38.9	31.4	32.7
Qwen-VL [7]	Qwen-7B	57.5	78.2	56.6	61.5	68.2	38.9	120.2	81.0	1487.5	60.6	-	58.2
LLaVA-v1.5 [48]	Vicuna-7B	62.0	78.5	-	58.2	66.8	50.0	-	-	1510.7	64.3	30.5	58.6
multimodal generalist LLM													
ImageBind-LLM [31]	LLaMA-7B	41.1	-	-	24.0	51.4	-	29.6	23.5	775.7	-	-	-
ChatBridge-13B [104]	Vicuna-13B	<u>41.8</u>	-	<u>45.2</u>	-	-	-	<u>115.7</u>	**82.5**	-	-	-	-
AnyMAL-13B [59]	LLaMA2-13B	-	59.6	33.1	24.7	52.7	24.4	-	-	-	-	-	-
AnyMAL-70B [59]	LLaMA2-70B	-	<u>64.2</u>	42.6	<u>32.9</u>	**70.8**	<u>33.8</u>	-	-	-	-	-	-
OneLLM-7B (Ours)	**LLaMA2-7B**	**59.5**	**71.6**	**58.9**	**34.0**	<u>63.4</u>	**45.9**	**115.9**	<u>78.6</u>	**1392.0**	**60.0**	**29.1**	**61.2**

Table 1. **Evaluation on 12 Image-Text Benchmarks**, including 6 VQA tasks (GQA [34], VQAv2 [26], OKVQA [55], TextVQA (TVQA) [75], ScienceQA (SQA) [52] and Vizwiz [29]), 2 image captioning tasks (Nocaps [2] and Flickr30K [65]), and 4 multimodal benchmarks (MME [20], MM Bench (MMB) [50], MMVet [98] and SEED [41]). The LLMs are Chinchilla [32], Vicuna [81], Qwen [6], LLaMA [78] and LLaMA2 [79]. The evaluation metrics for VQA and captioning tasks are accuracy and CIDEr, respectively. The results in **bold** and <u>underline</u> are the best and second-best results, respectively. -: Not reported result.

Model	0-shot	NextQA Acc.	How2QA Acc.	MSVD Acc.	VATEX CIDEr
HGQA [87]	✗	51.8	-	41.2	-
JustAsk [93]	✗	52.3	84.4	46.3	-
VALOR [12]	✗	-	-	60.0	95.1
SeViLA [97]	✗	73.8	83.6	-	-
FrozenBiLM [94]	✓	-	58.4	33.8	-
InternVideo [84]	✓	<u>49.1</u>	<u>62.2</u>	<u>55.5</u>	-
ChatBridge-13B [104]	✓	-	-	45.3	**48.9**
AnyMAL-13B [59]	✓	47.9	59.6	-	-
OneLLM-7B (Ours)	✓	**57.3**	**65.7**	**56.5**	<u>43.8</u>

Table 2. **Evaluation on Video-Text Tasks**, including video question answering (NextQA [86], How2QA [46] and MSVD [89]) and video captioning tasks (VATEX [83]). Acc.: Accuracy.

Model	0-shot	MUSIC-AVQA Acc.	VALOR CIDEr	AVSD CIDEr
MAVQA [42]	✗	71.5	-	-
VALOR [12]	✗	78.9	61.5	-
VAST [13]	✗	80.7	62.2	-
FA+HRED [61]	✗	-	-	84.3
MTN [40]	✗	-	-	98.5
COST [64]	✗	-	-	108.5
ChatBridge-13B [104]	✓	<u>43.0</u>	<u>24.7</u>	**75.4**
OneLLM-7B (Ours)	✓	**47.6**	**29.2**	<u>74.5</u>

Table 4. **Evaluation on Audio-Video-Text Tasks**, including audio-visual question answering on MUSIC-AVQA [42] and audio-visual captioning on VALOR-32K [12] and dialog completion on AVSD [3].

Model	0-shot	Clotho Caption		Clotho AQA
		CIDEr	SPIDEr	Acc.
FeatureCut [96]	✗	43.6	27.9	-
Wavcaps [56]	✗	48.8	31.0	-
MWAFM [43]	✗	-	-	22.2
Pengi [17]	✗	-	27.1	64.5
LTU-7B [25]	✓	-		11.9
ChatBridge-13B [104]	✓	26.2	-	-
OneLLM-7B (Ours)	✓	**29.1**	**19.5**	**57.9**

Table 3. **Evaluation on Audio-Text Tasks**, including audio captioning on Clotho Caption [18] and audio question answering on Clotho AQA [47].

Model	Captioning			Classification
	BLEU-1	ROUGE-L	METEOR	GPT4-Acc.
InstructBLIP-7B [15]	11.2	13.9	14.9	38.5
InstructBLIP-13B [15]	12.6	15.0	16.0	35.5
PointLLM-7B [92]	8.0	11.1	15.2	**47.5**
PointLLM-13B [92]	9.7	12.8	15.3	45.0
One-LLM-7B (Ours)	**42.2**	**45.3**	**20.3**	44.5

Table 5. **Evaluation on Point Cloud-Text Tasks**. The evaluation dataset is from Objaverse [16], following the data split in PointLLM [92]. InstructBLIP takes single-view image as input, while PointLLM and OneLLM take point cloud as input. GPT4-Acc.: GPT4 as the accuracy evaluator [92].

7B outperforms other MMLLMs such as ChatBridge-13B [104] and AnyMAL-13B [59] by a large margin. Our 7B model is even better than AnyMAL with 70B parameters. For image captioning tasks, OneLLM-7B is on-par with ChatBridge-13B. Although OneLLM is not specifically designed for vision tasks, our results demonstrate that

OneLLM can also reach the leading level in vision specialized LLMs, and the gap between MMLLMs and vision LLMs has further narrowed.

Video-Text Evaluation. As shown in Tab. 2, we evaluate OneLLM on video QA and captioning tasks. Our

Model	0-shot	NYUv2 Acc.	SUN RGB-D Acc.
ImageBind [23]	✗	54.0	35.1
Omnivore [22]	✗	76.7	64.9
Random	✓	10.0	5.26
CLIP ViT-H* [67]	✓	41.9	25.4
OneLLM-N (Ours)	✓	46.5	21.2
OneLLM-D (Ours)	✓	**50.9**	**29.0**

Table 6. **Evaluation on Scene Classification Tasks Using Depth / Normal Map.** OneLLM-N/D: OneLLM with Depth / Normal map inputs. Note that NYUv2 [60] and SUN RGB-D [76] only have depth maps, we adopt pretrained DPT model [19] to generate normal maps. *: The input to CLIP is depth rendered grayscale image. ImageBind is trained on image-depth pairs of SUN RGB-D and therefore is not zero-shot.

Task	NoCaps	VQAv2	ClothoQA	MSVDQA
(a) Training Mode				
Separate	115.6(-0.2)	71.9(+0.3)	37.8(-19.6)	31.0(-25.8)
Joint	115.8	71.6	57.4	56.8
(b) Weight Initialization				
Random Init.	98.8(-17.0)	65.6(-6.0)	57.6(+0.2)	53.1(-3.7)
Image Init.	115.8	71.6	57.4	56.8
(c) Number of Experts (Parameters)				
1 (88M)	108.7(-7.1)	66.9(-4.7)	58.2(+0.8)	53.3(-3.5)
3 (264M)	115.8	71.6	57.4	56.8
5 (440M)	114.6	71.7	58.2	56.7
7 (616M)	114.9	71.6	58.8	56.0
(d) Router Type				
Constant Router	109.8(-6.0)	67.7(-3.9)	56.2(-1.2)	55.3(-1.5)
Sparse Router	112.8(-3.0)	71.1(-0.5)	56.7(-0.7)	55.7(-1.1)
Soft Router	115.8	71.6	57.4	56.8

Table 7. **Ablation Experiments.** We choose three modalities (image, audio, video) and four datasets (NoCaps [2], VQAv2 [26], ClothoQA [47] and MSVDQA [89]) for evaluation. The row with gray background is our default setting.

model outperforms both MLLMs (ChatBridge and Any-MAL) and video-specific models (FrozenBiLM [94] and InternVideo [84]) in video QA tasks. Notably, our training datasets do not include video QA data like NextQA [86] and How2QA [46], which are video QA tasks that provide answer options. However, our model's training on similar VQA datasets (*e.g.*, A-OKVQA [71]) has evidently enhanced its emergent cross-modal capabilities, contributing to the improved performance in video QA tasks.

Audio-Text Evaluation. We evaluate OnLLM on audio captioning and QA tasks. In Tab. 3, we outperforms both ChatBridge and LTU [25] on Clotho Caption [18]. Notably, our zero-shot result on Clotho AQA [47] is on par with fully finetuned Pengi [17]. Similar to our conclusion on video QA, we believe that the captioning task requires more dataset-specific training, while the QA task may be a more accurate measure of the model's inherent zero-shot understanding capabilities.

Audio-Video-Text Evaluation. We evaluate OneLLM on audio-video-text tasks, such as QA (MUSIC AVQA [42]), captioning (VALOR-32K [12]) and dialog completion (AVSD [3]) based on the video and background audio. As shown in Tab. 4, OneLLM-7B surpasses ChatBridge-13B on all three datasets. Note that ChatBridge was trained on an audio-visual dataset [12], while OneLLM has not been trained on any audio-visual datasets. Since all modalities in OneLLM are well aligned with language, we can directly input video and audio signals to OneLLM during inference.

Point Cloud-Text Evaluation. In Tab. 5, We evaluate OneLLM on point cloud captioning and classification tasks. OneLLM can achieve excellent captioning results due to our carefully designed instruction prompts for switching between tasks (Sec. 3.3), while InstructBLIP [15] and PointLLM [92] struggle to generate short and accurate captions. On the classification task, OneLLM can also achieve comparable results to PointLLM.

Depth/Normal Map-Text Evaluation. Since there are currently no QA and captioning tasks using depth/normal maps, we evaluate OneLLM on two scene classification datasets [60, 76]. The performance, as displayed in Tab. 6, reveals that OneLLM achieves superior zero-shot classification accuracy compared to CLIP. These results affirm that OneLLM trained on synthetic depth/normal map data can adapt to real world scenarios.

IMU-Text and fMRI-Text Evaluation. Since IMU/fMRI to text generation are seldom explored in previous literature, we solely report our results on IMU/fMRI captioning. For IMU captioning on Ego4D [27], we evaluate OneLLM on a held-out subset with 2000 items. The CIDEr and ROUGE-L score are 24.9 and 19.5, respectively. For fMRI captioning on NSD [5], we evaluate OneLLM on its testing set, where OneLLM achieves 31.7 CIDEr and 25.1 ROUGE-L.

4.3. Ablation Experiments

In this section, we will explore some key designs of OneLLM. Our ablation experiments are conducted on a subset of the training data, which only includes multimodal alignment and instruction tuning datasets of image, audio and video, except for studies on the number of experts. Other settings remain unchanged if not specified.

Separate Training *vs.* Joint Training. An important question for MLLMs is whether a jointly trained MLLM is better than modality-specific MLLM? To address this, we compare the performance of separately trained MLLMs against a jointly trained MLLM in Tab. 7 (a). In separate training, the model can only access its own data; in joint training, the model is jointly trained on all data. On two image-text tasks NoCaps and VQAv2, we can see that separately and jointly trained models achieve comparable results; While
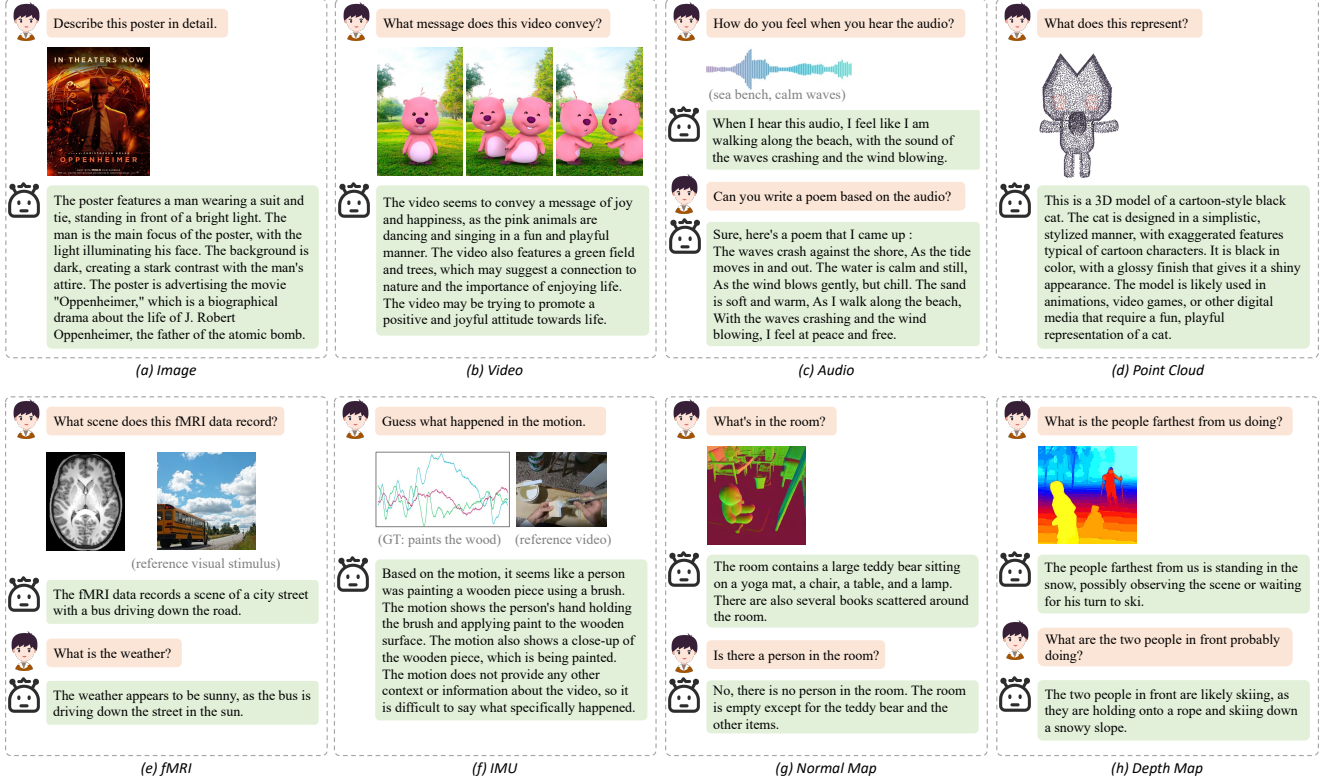
Figure 3. **Qualitative Results on Eight Modalities.** All demo inputs are from the web or the testing set of corresponding modalities.

separately trained audio and video models are much worse than the jointly trained model on ClothoQA and MSVDQA, respectively. This suggest that joint training substantially benefits data-scarce modalities (*e.g.*, audio and video), by allowing for the transfer of learned knowledge (*e.g.*, question answering) across modalities.

Image Alignment Benefits Multimodal Alignment. Tab. 7 (b) demonstrate that OneLLM with image-text alignment can help multimodal-text alignment. If we directly align all modalities with text using a random initialized model (*i.e.* universal projection module), the performance on image and video will drop significantly. Instead, OneLLM with image-text pretraining can better balance different modalities.

Number of Projection Experts. The number of projection experts in UPM is closely related to the number of modalities that OneLLM can accommodate. As shown in Tab. 7, OneLLM with three projection experts is enough to hold all modalities. Increasing the number of experts does not bring about the desired improvement, while the results with one expert is also not satisfactory.

Router Type. The modality router is to link multiple projection experts into a single module. Here we discuss three types of router: constant router, sparse router and the default soft router. **(a)** Constant router links K experts with a constant number $1/K$. The output of constant router is $\sum_{k=1}^{K} \frac{1}{K} \cdot P_k(\mathbf{x})$. **(b)** Sparse router only selects one expert with the maximum routing weight. The output is $w_{k^*} P_{k^*}(\mathbf{x})$ where $k^* = \arg\max_k w_k$. As shown in Tab. 7 (d), soft router outperforms other two routers, indicating its effectiveness for dynamic routing of multimodal signals.

4.4. Qualitative Analysis

Fig. 3 gives some qualitative results of OneLLM on eight modalities. We show OneLLM can **(a)** understand both visual and textual content in images, **(b)** leverage temporal information in videos, **(c)** do creative writing based on audio content, **(d)** understand the details of 3D shapes, **(e)** analyze visual scenes recorded in fMRI data, **(f)** guess the person's action based on motion data, and **(g)-(h)** scene understanding using depth/normal map. Due to space limit, we put more qualitative results to Sec. F of the appendix.

5. Conclusion

In this work, we introduce OneLLM, an MLLM that aligns eight modalities with language using a unified framework. Initially, we first train a basic vision LLM. Building on this, we design a multimodal framework with a universal encoder, a UPM and an LLM. By a progressive alignment pipeline, OneLLM can handle multimodal inputs with a single model. Furthermore, we curate a large-scale multimodal

instruction dataset to fully unleash OneLLM's instruction-following capability. Finally, we evaluate OneLLM on 25 diverse benchmarks, showing its excellent performance.

Limitation and Future Work. Our work faces two primary challenges: **(i)** The absence of large-scale, high-quality datasets for modalities beyond image, which leads to a certain gap between OneLLM and specialized models on these modalities. **(ii)** Fine-grained multimodal understanding in high-resolution images, long sequences video and audio *etc.* In the future, we will collect high-quality datasets and design new encoders to realize fine-grained multimodal understanding, *e.g.*, supporting varying length inputs [9].

A. Appendix Overview

B. Additional Ablation Experiments

Encoder Type	Frozen	Mem.	Nocaps	VQAv2	ClothoQA	MSVDQA
CLIP	✓	46Gb	115.8	71.6	57.4	56.8
CLIP	✗	74Gb	106.0(-9.8)	69.1(-2.5)	62.1(+4.7)	53.6(-3.2)
DINOv2	✓	33Gb	104.6(-11.2)	67.0(-4.6)	56.8(-0.6)	54.7(-2.1)

Table 8. **Ablation Experiments on Universal Encoder.**

In the main paper, we follow previous works [103] and set a frozen CLIP-ViT as the universal encoder. Here we explore other design choices such as trainable CLIP-ViT and DINOv2 [63] as the encoder.

Frozen *vs.* Trainable Encoder. We first turn on all the parameters in the multimodal-text alignment stage. As shown in Tab. 8, the performance for visual modalities (image and video) dropped significantly, while the result for audio QA (ClothoQA) improved by 4.7%. We think trainable CLIP will break the pretrained vision-language representations but can leave more space for learning other modalities. However, considering the memory usage (46Gb *vs.* 74Gb), frozen CLIP will be a better choice for our framework.

Beyond Vision-Language Encoder. In addition to the vision-language encoder CLIP-ViT, we also explore other models, such as the self-supervised vision model DI-NOv2 [63], as the universal encoder. In Tab. 8, we noticed that the performance of OneLLM using DINOv2 is lower than the model using CLIP-ViT because DINOv2 is not aligned with language and we need to learn the vision-language alignment from scratch.

C. Additional Implementation Details

C.1. Lightweight Modality Tokenizers

The modality tokenizer is to transform input signal into a sequence of tokens. Here we will introduce the tokenizer of each modality in detail.

Visual Tokenizer. We use the same tokenizer setting for visual modalities, *i.e.*, image, video, depth/normal map. The visual tokenizer is a single 2D convolution layer:

$$\text{Conv2D}(C_{in} = 3, C_{out} = 1024, K = (14, 14), S = (14, 14)), \quad (3)$$

where C_{in}, C_{out}, K and S denote the input channel, output channel, kernel size and stride, respectively. Note that for a video input $\mathbf{x} \in \mathbb{R}^{T \times H \times W}$ with T frames, height H and width W, we parallel feed its frames into the tokenizer, resulting in $T \times \frac{H}{14} \times \frac{W}{14}$ tokens. Similarly, image, depth/normal map can also be regarded as a one-frame video input $\mathbf{x} \in \mathbb{R}^{1 \times H \times W}$.

Audio Tokenizer. We first transform audio signals into 2D spectrogram features $\mathbf{x} \in \mathbb{R}^{1 \times H \times W}$, where H=128 and W=1024 by default. Following [24], the audio tokenizier is a single 2D convolution layer:

$$\text{Conv2D}(C_{in} = 1, C_{out} = 1024, K = (16, 16), S = (10, 10)). \quad (4)$$

Point Tokenizer. For a raw point cloud, we sample 8192 points using Furthest Point Sampling (FPS), resulting in a 2D tensor $\mathbf{x} \in \mathbb{R}^{8192 \times 6}$. Then we use the KNN algorithm to group these points into 512 groups: $\mathbf{x} \in \mathbb{R}^{512 \times 32 \times 6}$ where 32 is the size of each group. After that, we encode the point cloud with a 2D convolution layer:

$$\text{Conv2D}(C_{in} = 6, C_{out} = 1024, K = (1, 1), S = (1, 1)), \quad (5)$$

followed by a `max` operation on dimension 1. Finally, the shape of output tokens is $\mathbb{R}^{1024 \times 1024}$.

IMU Tokenizer. For an IMU input with shape $\mathbb{R}^{2000 \times 6}$, we tokenize it with a 1D convolution layer:

$$\text{Conv1D}(C_{in} = 6, C_{out} = 1024, K = 10, S = 1), \quad (6)$$

resulting in a sequence of tokens $\mathbf{x} \in \mathbb{R}^{1024 \times 391}$.

fMRI Tokenizer. The shape of an fMRI signal is $\mathbb{R}^{15724}$. We tokenize it with a 1D convolution layer:

$$\text{Conv1D}(C_{in} = 15724, C_{out} = 8196, K = 1, S = 1). \quad (7)$$

We then resize the output tensor $\mathbf{x} \in \mathbb{R}^{8196}$ into a 2D tensor $\mathbf{x} \in \mathbb{R}^{1024 \times 8}$ to align with the input of the transformer encoder.

C.2. Multimodal-Text Alignment Dataset

We summary the multimodal-text alignment dataset in Tab. 9. For depth/normal-text pairs, we adopt DPT model [68] pretrained on ominidata [19] to generate depth/normal map. The source dataset is a subset of CC3M [73], around 0.5M image-text pairs. For IMU-text pairs, we use the IMU sensor data of Ego4D [27] and the corresponding video narrations (*i.e.*, text annotations). For fMRI-text pairs, we use the subj01 imaging session of NSD [5] and follow the same data split with [72]. Note that the visual stimulus, *i.e.*, images shown to participants, are from MS COCO [14]. Therefore, we use the image captions in COCO Captions as text annotations of fMRI-text pairs.

Modality	Multimodal-Text Alignment		Multimodal Instruction Tuning	
	Size	Dataset	Size	Dataset
Image	1000M	LAION-400M [70] LAION-COCO [69]	1216K	LLaVA-150K [49] COCO Caption [14] VQAv2 [26], GQA [34] OKVQA [55], A-OKVQA [71] OCRVQA [58], RefCOCO [36] Visual Genome [38]
Video	2.5M	WebVid-2.5M [8]	461K	MSRVTT-Cap [91] MSRVTT-QA [89] Video Conversation [104]
Audio	0.4M	WavCaps [56]	60K	AudioCaps [37] Audio Conversation [104]
Point	0.6M	Cap3D [54]	70K	Point Conversation [92]
Depth	0.5M	CC3M [73]	50K	LLaVA-150K [49]
Normal	0.5M	CC3M [73]	50K	LLaVA-150K [49]
IMU	0.5M	Ego4D [27]	50K	Ego4D [27]
fMRI	9K	NSD [5]	9K	NSD [5]
Text	-	-	40K	ShareGPT [1]
Total	**1005M**		**2006K**	-

Table 9. **Training Datasets.**

C.3. Multimodal Instruction Tuning Dataset

We summary the multimodal instruction tuning dataset in Tab. 9.

C.4. Prompt Design

The prompt formats for each dataset are shown in Tab. 10.

D. Evaluation Details

In this section, we first list the evaluation prompts for each dataset in Tab. 11. Then we will give more evaluation details.

Image, Video and Audio Tasks. We evaluate all datasets using their official evaluation protocols. As shown in Tab. 11, for QA tasks with options, we ask OneLLM to directly predict the option letters; For open-ended QA tasks,

Dataset	Prompt Format
LLaVA-150K [49] ShareGPT [1] Video Conversation [104] Audio Conversation [104] Point Conversation [92]	(use their original prompt)
VQAv2 [26], GQA [34] OKVQA [71] OCRVQA [58] MSRVTT-QA [89]	{Question} Answer the question using a single word or phrase.
A-OKVQA [71]	{Question} {Options} Answer with the option's letter from the given choices directly
TextCaps [74] COCO Caption [14] MSRVTT-Cap [91] AudioCaps [37]	Provide a one-sentence caption for the provided image/video/audio.
RefCOCO [36] Visual Genome [38]	Provide a short description for this region.
Ego4D [27]	Describe the motion.
NSD [5]	Describe the scene based on fMRI data.

Table 10. **Prompt Formats for Training.**

Dataset	Prompt Format
MMVet	(use the original prompt)
GQA [34] VQAv2 [26] OKVQA [55] TextVQA [75] MME [20] MSVD [89] Clotho AQA [47] MUSIC-AVQA [42]	{Question} Answer the question using a single word or phrase.
ScienceQA [52] MMbench [50] SEED-Bench [41] NextQA [86] How2QA [46]	{Question} {Options} Answer with the option's letter from the given choices directly
VizWiz [29]	{Question} When the provided information is insufficient, respond with 'Unanswerable'. Answer the question using a single word or phrase.
Nocaps [2] Flickr30K [65] VATEX [83] VALOR [12] Clotho Cap [18] Objaverse-Cap [16]	Provide a one-sentence caption for the provided image/video/audio/point cloud.
AVSD [3]	{Question} Answer the question and explain the reason in one sentence.
Objaverse-CLS [16]	What is this?
NYUv2 [60] SUN RGB-D [76]	{Class List} What is the category of this scene? Choice one class from the class sets.

Table 11. **Prompt Formats for Evaluation.**

we ask OneLLM to predict a single word or phase. For captioning tasks, we ask OneLLM to generate a one-sentence caption. Note that for audio-video-text tasks, the input sequence to the LLM is: {*Video Tokens*} {*Audio Tokens*} {*Text Prompts*}.

Model	Encoder Param	#Encoder	#Projection	Image	Video	Audio	Point	IMU	Depth	Normal	fMRI
X-LLM [10]	-	3	3	✓	✓	✓					
PandaGPT [77]	1.2B	2	1	✓	✓	✓					
ImageBind-LLM [31]	1.8B	3	1	✓	✓	✓	✓				
ChatBridge [104]	1.3B	3	3	✓	✓	✓					
AnyMAL [59]	2B	3	3	✓	✓	✓		✓			
OneLLM (Ours)	**0.6B**	**1**	**1**	✓	✓	✓	✓	✓	✓	✓	✓

Table 12. **Comparisons of Different Multimodal LLMs.**

Point Cloud Tasks. Our evaluation on point cloud tasks mainly follows PointLLM [92]. For the point cloud classification task, we use the same prompt as PointLLM: *What is this*, and evaluate the accuracy using GPT4.

Depth/Normal Map Tasks. For scene classification using depth/normal map, we first prepend the category list to the beginning of prompt, then we ask OneLLM to choose one class for the list.

IMU/fMRI Tasks. We evaluate on IMU/fMRI captioning tasks. The prompts are the same as their training prompts: *Describe the motion* for IMU captioning and *Describe the scene based on fMRI data* for fMRI captioning.

E. Comparison with Prior Works

The main difference between OneLLM and previous MLLMs is that we show a unified encoder is *sufficient* to align multi-modalities with LLMs. As shown in Tab. 12, OneLLM with *one* universal encoder, *one* projection module and *less* parameters (0.6B) can unify more modalities into one framework. The results in the main paper (Tab.1-6) also demonstrate that OneLLM can achieves better performance to previous works. The ablation experiments in Tab.7 (a) also show that jointly training all modalities with our unified framework can benefit data-scarce modalities. Here we are not trying to prove that OneLLM's architecture is optimal, but to show the possibility of building MLLMs using a unified and scalable framework.

F. Additional Qualitative Results

In this section, we provide more qualitative results in Fig. 4, Fig. 5 and Fig. 6.

References

[1] Sharegpt. https://sharegpt.com/, 2023. 10

[2] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. nocaps: novel object captioning at scale. In *ICCV*, 2019. 6, 7, 10

[3] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K. Marks, Chiori Hori, Peter Anderson, Stefan Lee, and Devi Parikh. Audio-visual scene-aware dialog. In *CVPR*, 2019. 6, 7, 10

[4] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 35:23716–23736, 2022. 1, 2, 6

[5] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022. 5, 7, 10

[6] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 6

[7] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 6

[8] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, pages 1728–1738, 2021. 5, 10

[9] Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, Arushi Somani, and Sağnak Taşırlar. Introducing our multimodal models, 2023. 9

[10] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages. *arXiv preprint arXiv:2305.04160*, 2023. 2, 3, 11

[11] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 1, 3

[12] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and

What do you think of the photo?

The photo is a captivating and dynamic scene of a bustling city street in New York City. The image features a large number of billboards and advertisements on the buildings surrounding the street, creating a vibrant atmosphere. The street is filled with people walking and cars driving, giving the impression of a busy urban environment. The photo is taken from a low angle, which adds a sense of depth and perspective to the scene. The traffic lights are visible, indicating that the street is well-regulated for pedestrian and vehicular safety. The overall composition of the photo is visually striking, showcasing the energy and excitement of the city.

Where might this picture have been taken?

This picture might have been taken in a large city with a lot of tall buildings. The presence of a red double-decker bus and a truck in the image suggests that the city has a public transportation system and possibly a busy traffic environment. The tall buildings and the presence of a bus and truck indicate that the city is likely a bustling urban area with a high population density.

Choose one from "Sydney", "New York" and "Hong Kong", and tell me the reason.

The presence of a red double-decker bus and a truck in the image suggests that the city could be either Hong Kong or New York.

Why this image is unusual?

The image is unusual because it features a dog dressed in a chef's outfit, complete with a hat, apron, and a bow tie. This is not a typical sight, as dogs are not typically dressed in human clothing or accessories. The dog is sitting at a table with a bowl of food, which adds to the unconventional nature of the scene. The image is likely meant to be humorous or creative, as it is not a common occurrence to see a dog dressed in human attire.
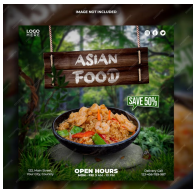
Tell me about this place.

The image features a large, grassy field with a group of stone structures, possibly a stonehenge, situated in the center. The structures are made of rocks and are arranged in a circle, with some rocks appearing to be taller than others. The sky above the field is filled with clouds, creating a beautiful and serene atmosphere. The scene is reminiscent of ancient civilizations and their architectural achievements.

What does this picture want to express?

The image is a symbolic representation of the bond between a parent and their child. The hand of an adult is holding a small plant, which is a symbol of growth and life, while the hand of a baby is reaching out to the plant. This image suggests the nurturing and care that parents provide to their children, as well as the innocence and curiosity of a child. The scene is set in a grassy field, which adds to the natural and serene atmosphere of the image.

Introduce me this poster.

The poster features a delicious Asian food advertisement, with a large bowl of rice and shrimp displayed prominently. The advertisement is designed to entice customers with its mouth-watering presentation. The poster is displayed in a restaurant, with a sign that reads "Open Hours" and "50% Off" in the background. This suggests that the restaurant is offering a special deal on Asian food during certain hours.

Why this image is funny?

This image is funny because it features three raccoons dressed in human clothes and driving a car. The raccoons are wearing sunglasses and one of them is playing a guitar, which adds a humorous and unexpected element to the scene. The fact that raccoons are not capable of driving or playing instruments makes the situation comical and amusing to viewers.

If you could only use three words to express the theme of this picture, what would they be?

Foggy, tree, person

Provide a short caption for this image

A person running in a field with a tree in the background.

What iconic attractions appear in this picture?

The picture features several iconic attractions, including the Statue of Liberty, the Eiffel Tower, the Roman Coliseum, the pyramids of Egypt, the Space Needle, and the London Bridge.
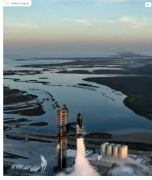
Figure 4. **Additional Qualitative Image Demos.**

Describe the video in detail

A man is playing basketball in a gym. He is wearing a black shirt and white pants. He is bouncing a basketball on his finger and then dribbling it with his hand. He is also wearing a black and white striped shirt. The gym is white and has a white floor. There is a white wall behind him.

What activity is going on in the video?

The video shows a rocket launching from a launch pad.

Where might this launch pad be located?

It is not clear from the video where the launch pad is located. It could be in a spaceport or a military base.

What process does this video record?

This video records the process of a flower blooming.

Can you provide more details?

The flower is white and has a long stem. The flower is in a greenhouse, and the camera is recording the flower's growth process. The flower is in bloom, and the camera captures the flower's growth and development.

What can you hear in the audio?

(bell ringing)

I can hear the sound of a bell ringing at a steady pace.

Can you provide more details?

The bell is large and metallic, and it is ringing in a melody that is loud and clear.

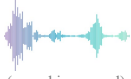**Where could this audio take place?**

(bird singing)

The audio could take place in a place with birds, like a forest or a park.

What kind of bird call could this be?

It could be a bird call like a whistle or a chirp, but without more context, it's hard to say for sure.

Describe the audio in detail

(crunching sound)

In the audio, a person is seen shuffling through a variety of items, and the items make a crunching sound. The person then takes a moment to speak. The items in the background are not clear, but they appear to be the source of the crunching sound.

Summary the content of the audio briefly.

(speaking and laughing)

A woman speaking followed by a baby laughing

Describe the audio in detail

In the audio, a woman is speaking to a laughing baby. The woman's voice is calm and soothing, while the baby's laughter is contagious. It sounds like they are having a conversation, with the woman saying something that makes the baby laugh even more. The audio also contains some background noise, but it is not clear what it is.

What is this?

This is a 3D model of a cartoon-style humanoid figure, predominantly white in color. The figure is designed in a simplistic, stylized manner, typical of cartoon characters. It has a neutral expression, suggesting it could be used in a variety of contexts, such as animation, game development, or graphic design. The figure's white color allows for easy customization, such as adding clothing or other accessories.

How would you summarize this 3D data set?

This 3D data set represents a toy car, predominantly blue in color, with a distinctive white stripe running along its length. The car features a black windscreen and matching black windows, adding to its realistic appearance. The toy car, with its vibrant colors and detailed design, is likely to be a popular choice among children for play and could also serve as a decorative item.

What kind of object is illustrated by this collection of points?
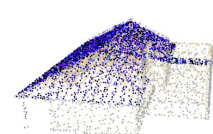
This is a 3D model of a cartoon-style dog, predominantly in a vibrant yellow color. The model is characterized by its exaggerated features, typical of cartoon aesthetics, such as large, expressive eyes and a broad, friendly smile. The dog's design suggests it could be used in animations, video games, or children's educational content.

Provide an outline of this 3D shape's characteristics.

The 3D object model is a tree with a distinctive green color. It features a sturdy trunk and a dense canopy of leaves, giving it a realistic appearance. The tree's design suggests it could be used in various digital environments, such as video games, animations, or virtual reality simulations.

Describe the 3D model in detail

The 3D model is a cartoon-style house with a vibrant purple roof. The house is simplistic in design, with a single story and a single room. The roof is a striking purple color, which contrasts with the white walls. The model is likely used in animation or game design, where such whimsical, colorful elements are common.

Figure 5. **Additional Qualitative Video, Audio and Point Cloud Demos.**

Describe the scene based on the normal map

The normal map shows a group of horses standing in a fenced-in area, possibly a corral. There are five horses in total, with four of them standing close together, and one horse standing slightly apart from the others. The horses are of various sizes and are positioned in a way that suggests they are waiting or resting. The fence surrounding the area is made of wood, and the ground is covered with dirt.

What's in this room?

The room is dimly lit, with a bed in the center and a small wooden table nearby. There is also a chair in the room. The bed is made up with a white comforter and pillows, and there is a lamp on the table. The room is decorated with several potted plants, including a large one near the bed. Additionally, there is a vase with flowers in it, and a clock on the wall. The overall atmosphere of the room is cozy and inviting.

What are people doing in the normal map?

In the normal map, people are skateboarding and performing tricks at a skate park.
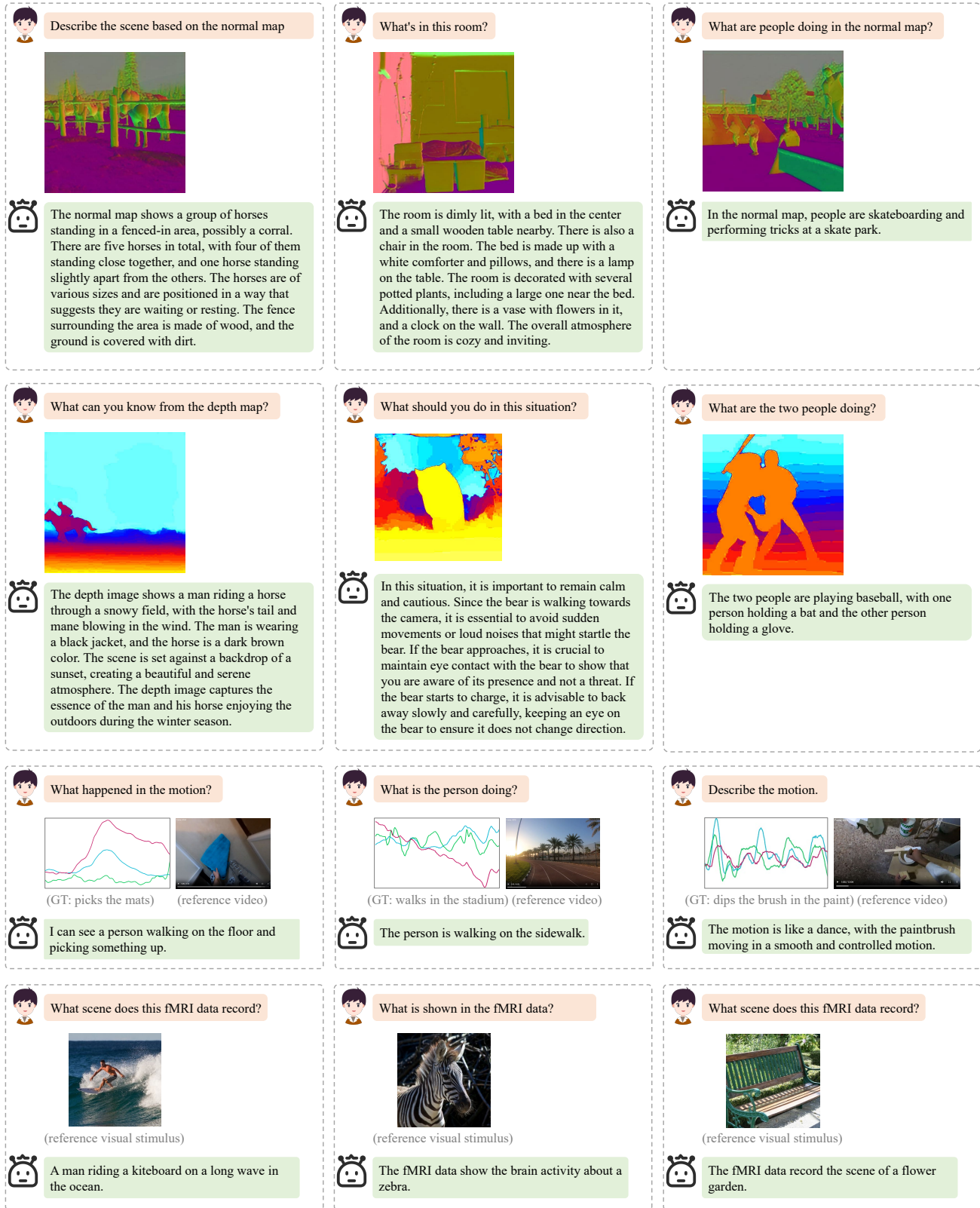
What can you know from the depth map?

The depth image shows a man riding a horse through a snowy field, with the horse's tail and mane blowing in the wind. The man is wearing a black jacket, and the horse is a dark brown color. The scene is set against a backdrop of a sunset, creating a beautiful and serene atmosphere. The depth image captures the essence of the man and his horse enjoying the outdoors during the winter season.

What should you do in this situation?

In this situation, it is important to remain calm and cautious. Since the bear is walking towards the camera, it is essential to avoid sudden movements or loud noises that might startle the bear. If the bear approaches, it is crucial to maintain eye contact with the bear to show that you are aware of its presence and not a threat. If the bear starts to charge, it is advisable to back away slowly and carefully, keeping an eye on the bear to ensure it does not change direction.

What are the two people doing?

The two people are playing baseball, with one person holding a bat and the other person holding a glove.

What happened in the motion?

(GT: picks the mats)   (reference video)

I can see a person walking on the floor and picking something up.

What is the person doing?

(GT: walks in the stadium) (reference video)

The person is walking on the sidewalk.

Describe the motion.

(GT: dips the brush in the paint) (reference video)

The motion is like a dance, with the paintbrush moving in a smooth and controlled motion.

What scene does this fMRI data record?

(reference visual stimulus)

A man riding a kiteboard on a long wave in the ocean.

What is shown in the fMRI data?

(reference visual stimulus)

The fMRI data show the brain activity about a zebra.

What scene does this fMRI data record?

(reference visual stimulus)

The fMRI data record the scene of a flower garden.

Figure 6. **Additional Qualitative Depth/Normal Map, IMU and fMRI Demos.**

dataset. *arXiv preprint arXiv:2304.08345*, 2023. 3, 6, 7, 10

[13] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv preprint arXiv:2305.18500*, 2023. 3, 6

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 5, 10

[15] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 5, 6, 7

[16] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023. 6, 10

[17] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks. *arXiv preprint arXiv:2305.11834*, 2023. 6, 7

[18] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, pages 736–740. IEEE, 2020. 6, 7, 10

[19] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, pages 10786–10796, 2021. 5, 7, 10

[20] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 6, 10

[21] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023. 2, 3, 4, 6

[22] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, pages 16102–16112, 2022. 7

[23] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, pages 15180–15190, 2023. 3, 7

[24] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575, 2021. 9

[25] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023. 1, 3, 6, 7

[26] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: El-

[27] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 5, 7, 10

[28] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 1, 3

[29] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*, pages 3608–3617, 2018. 6, 10

[30] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, pages 976–980. IEEE, 2022. 3

[31] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023. 2, 3, 6, 11

[32] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 6

[33] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[34] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 5, 6, 10

[35] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. pages 4651–4664. PMLR, 2021. 2, 3

[36] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798, 2014. 5, 10

[37] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 5, 10

[38] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123:32–73, 2017. 5, 10

[39] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang,

Siddharth Karamcheti, Alexander M Rush, Douwe Kiela, et al. Obelisc: An open web-scale filtered dataset of interleaved image-text documents. *arXiv preprint arXiv:2306.16527*, 2023. 6

[40] Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*, 2019. 6

[41] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023. 6, 10

[42] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, pages 19108–19118, 2022. 5, 6, 7, 10

[43] Guangyao Li, Yixin Xu, and Di Hu. Multi-scale attention for audio question answering. *arXiv preprint arXiv:2305.17993*, 2023. 6

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1, 2, 3, 6

[45] Kunchang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 1, 3

[46] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 6, 7, 10

[47] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. Clotho-aqa: A crowd-sourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE, 2022. 6, 7, 10

[48] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3, 5, 6

[49] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2023. 4, 5, 10

[50] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multimodal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 6, 10

[51] Kevin Lu, Aditya Grover, Pieter Abbeel, and Igor Mordatch. Pretrained transformers as universal computation engines. *arXiv preprint arXiv:2103.05247*, 1, 2021. 2, 4

[52] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *NeurIPS*, 2022. 6, 10

[53] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 3

[54] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. Scalable 3d captioning with pretrained models. *arXiv preprint arXiv:2306.07279*, 2023. 5, 10

[55] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *CVPR*, 2019. 5, 6, 10

[56] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023. 5, 6, 10

[57] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines. In *CoRL*, 2023. 2

[58] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, pages 947–952. IEEE, 2019. 5, 10

[59] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, et al. Anymal: An efficient and scalable any-modality augmented language model. *arXiv preprint arXiv:2309.16058*, 2023. 2, 3, 6, 11

[60] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 7, 10

[61] Dat Tien Nguyen, Shikhar Sharma, Hannes Schulz, and Layla El Asri. From film to video: Multi-turn question answering with multi-modal context. *arXiv preprint arXiv:1812.07023*, 2018. 6

[62] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 1

[63] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 9

[64] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video dialog as conversation about objects living in space-time. In *ECCV*, pages 710–726. Springer, 2022. 6

[65] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *ICCV*, pages 2641–2649, 2015. 6, 10

[66] Joan Puigcerver, Carlos Riquelme, Basil Mustafa, and Neil Houlsby. From sparse to soft mixtures of experts. *arXiv preprint arXiv:2308.00951*, 2023. 2, 4

[67] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language super-

vision. In *ICML*, pages 8748–8763. PMLR, 2021. 2, 3, 4, 7

[68] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021. 5, 10

[69] C. Schuhmann, A. Köpf, R. Vencu, T. Coombes, and R. Beaumont. Laion coco: 600m synthetic captions from laion2b-en. 5, 10

[70] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, 35: 25278–25294, 2022. 1, 5, 10

[71] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *ECCV*, pages 146–162. Springer, 2022. 5, 7, 10

[72] Paul S Scotti, Atmadeep Banerjee, Jimmie Goode, Stepan Shabalin, Alex Nguyen, Ethan Cohen, Aidan J Dempster, Nathalie Verlinde, Elad Yundler, David Weisberg, et al. Reconstructing the mind's eye: fmri-to-image with contrastive learning and diffusion priors. *arXiv preprint arXiv:2305.18274*, 2023. 10

[73] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018. 5, 10

[74] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, pages 742–758. Springer, 2020. 10

[75] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 6, 10

[76] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576, 2015. 7, 10

[77] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023. 3, 11

[78] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2, 6

[79] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 4, 5, 6

[80] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*, 2019. 4

[81] Vicuna. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. https://vicuna.lmsys.org/, 2023. 6

[82] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 3

[83] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4581–4591, 2019. 6, 10

[84] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6, 7

[85] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5. IEEE, 2023. 3

[86] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *CVPR*, pages 9777–9786, 2021. 6, 7, 10

[87] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *AAAI*, pages 2804–2812, 2022. 6

[88] Chenfeng Xu, Shijia Yang, Tomer Galanti, Bichen Wu, Xiangyu Yue, Bohan Zhai, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Image2point: 3d point-cloud understanding with 2d image pretrained models. In *ECCV*, pages 638–656. Springer, 2022. 2

[89] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 5, 6, 7, 10

[90] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 3

[91] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, pages 5288–5296, 2016. 5, 10

[92] Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. Pointllm: Empowering large language models to understand point clouds. *arXiv preprint arXiv:2308.16911*, 2023. 3, 5, 6, 7, 10, 11

[93] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, pages 1686–1697, 2021. 6

[94] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via

frozen bidirectional language models. *NeurIPS*, 35:124–141, 2022. 6, 7

[95] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 2, 3

[96] Zhongjie Ye, Yuqing Wang, Helin Wang, Dongchao Yang, and Yuexian Zou. Featurecut: An adaptive data augmentation for automated audio captioning. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 313–318. IEEE, 2022. 6

[97] Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *arXiv preprint arXiv:2305.06988*, 2023. 6

[98] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023. 6

[99] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023. 1, 3

[100] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. 1, 3

[101] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, pages 8552–8562, 2022. 3

[102] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023. 2

[103] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv preprint arXiv:2307.10802*, 2023. 2, 4, 9

[104] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. Chatbridge: Bridging modalities with large language model as a language catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 2, 3, 5, 6, 10, 11

[105] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3, 4